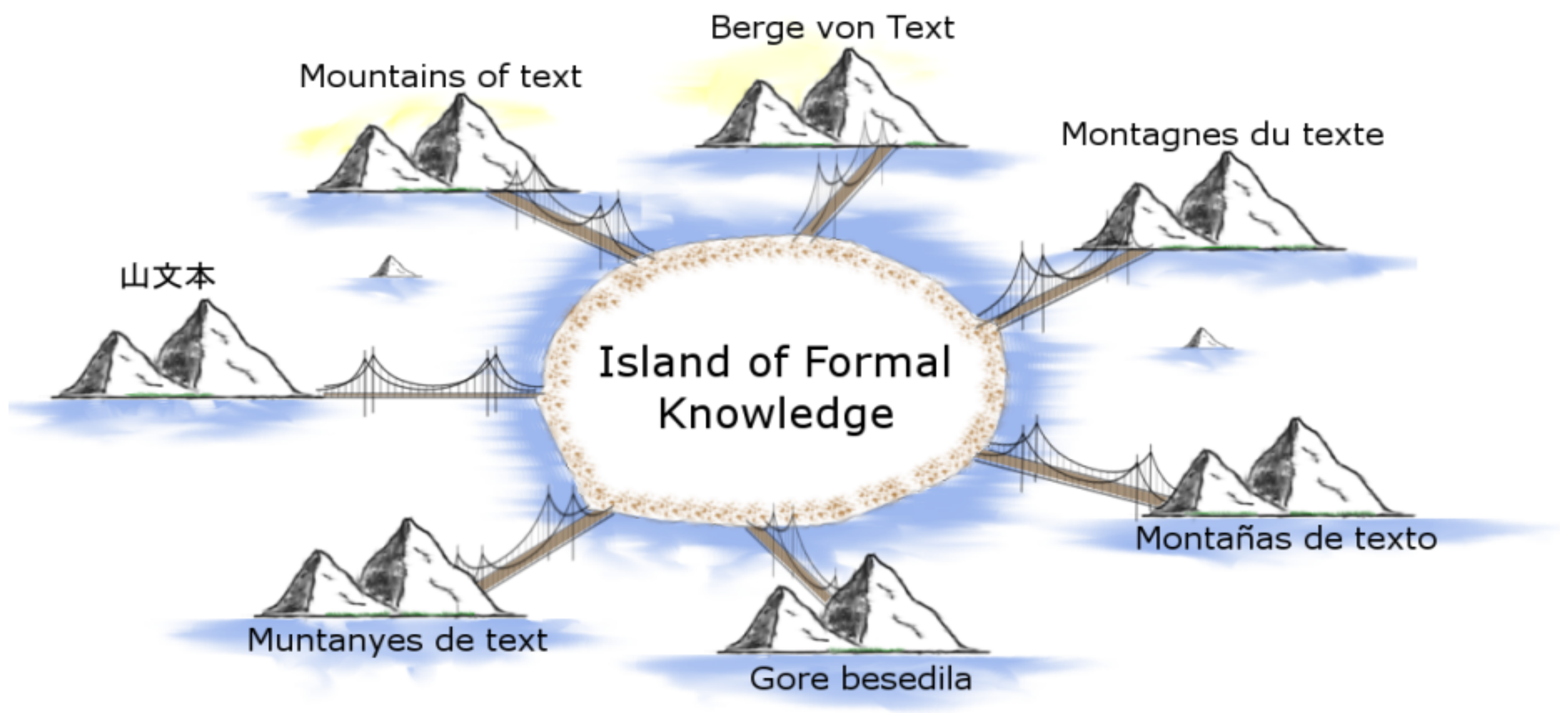
# XLike: Cross-lingual Knowledge Extraction

Marko Tadić
University of Zagreb, Faculty of Humanities and Social Sciences, marko.tadic@ffzg.hr

EAMT2014, Dubrovnik
2014-06-17

## Goal

The goal of the XLike project is to develop technology to monitor and aggregate knowledge that is currently spread across mainstream and social media, and to enable cross-lingual services for publishers, media monitoring and business intelligence.



## Research contributions

To extract and integrate formal knowledge from multilingual texts with cross-lingual knowledge bases and to adapt linguistic techniques and crowdsourcing to deal with irregularities in the informal language used primarily in social media.

### Languages covered:

a) Major languages: **English**, **German**, **Spanish**, **Chinese**
b) Minor languages: **Catalan**, **Slovenian** and **Croatian**

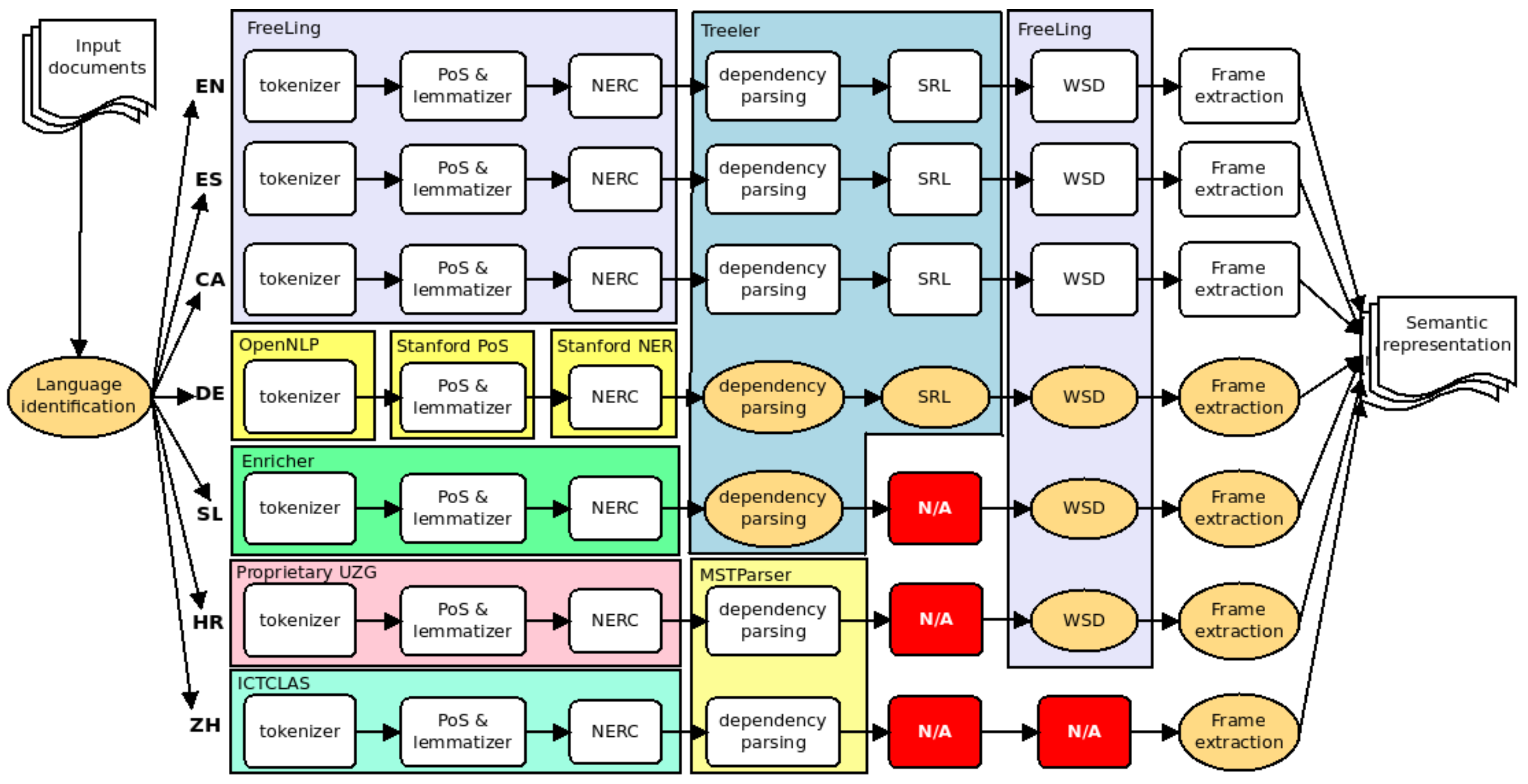
### Knowledge resources used as **interlingua**:

a) Linked Open Data (e.g. DBpedia)
b) Common sense knowledge base CycKB

For languages where no required linguistic resources are available, we will use a probabilistic Interlingua representation trained from a parallel corpora or comparable corpus derived from the Wikipedia.

## Linguistic processing

Fully automatized pipelines for tokenization, POS-tagging, lemmatization, NERC, dependency parsing, semantic role labelling for all seven main XLike languages.



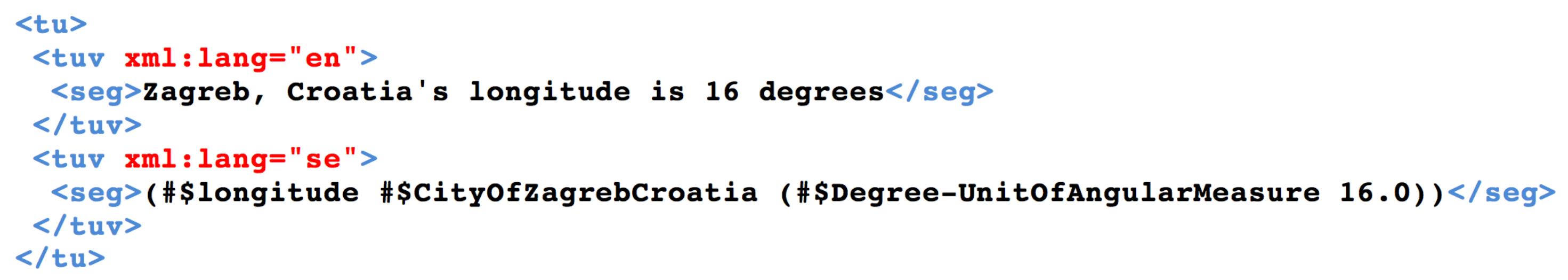
## MT in XLike

### Supporting technology in two cases:

a) translation from natural language (SL) to semantic representation in formal language (TL);
b) translation from under-resourced language(s) (SL) into English (TL) for processing with en-pipeline.

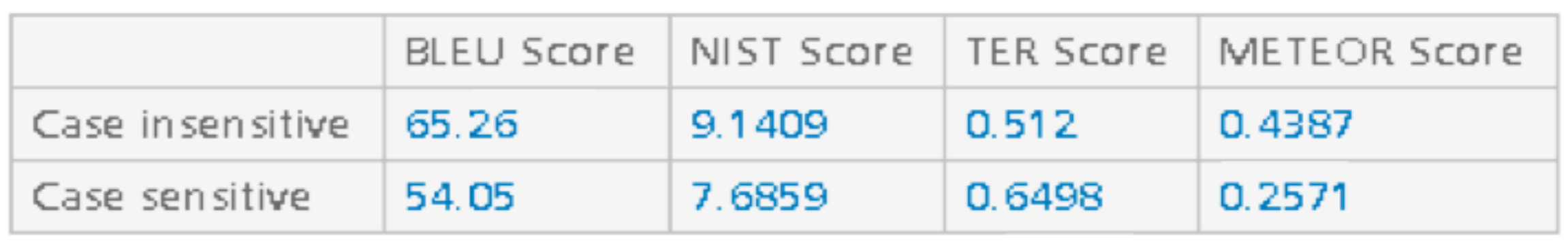
### The first attempt of SMT from natural language (NL, English) to formal language (FL, CycL)

- FL should be easier to generate
  - fixed word order: the notorious problem in SMT are TLs with free word order;
  - formal syntax: no syntactic irregularities that usually appear in NL texts, no phrases in TL that have to be treated as single units;
  - no NL morphology: often errors in inflectional endings contribute to lower fluency of TL
- CycL as FL
  - concepts and predicates are constants (prefixed by **#$**)
  - **#$isa** predicate: **(#$isa #$BarackObama #$UnitedStatesPresident)**
  - **#$genls** predicate: **(#$genls #$BabyOil #$BabyToiletrySubstance)**
  - **#$capitalCity** predicate: **(#$capitalCity #$Croatia #$Zagreb)**
- Using parallel corpus English-CycL
  - generated English sentences our of Cyc Ontology
  - 650,000 aligned “sentences”, 10,000 aside for evaluation

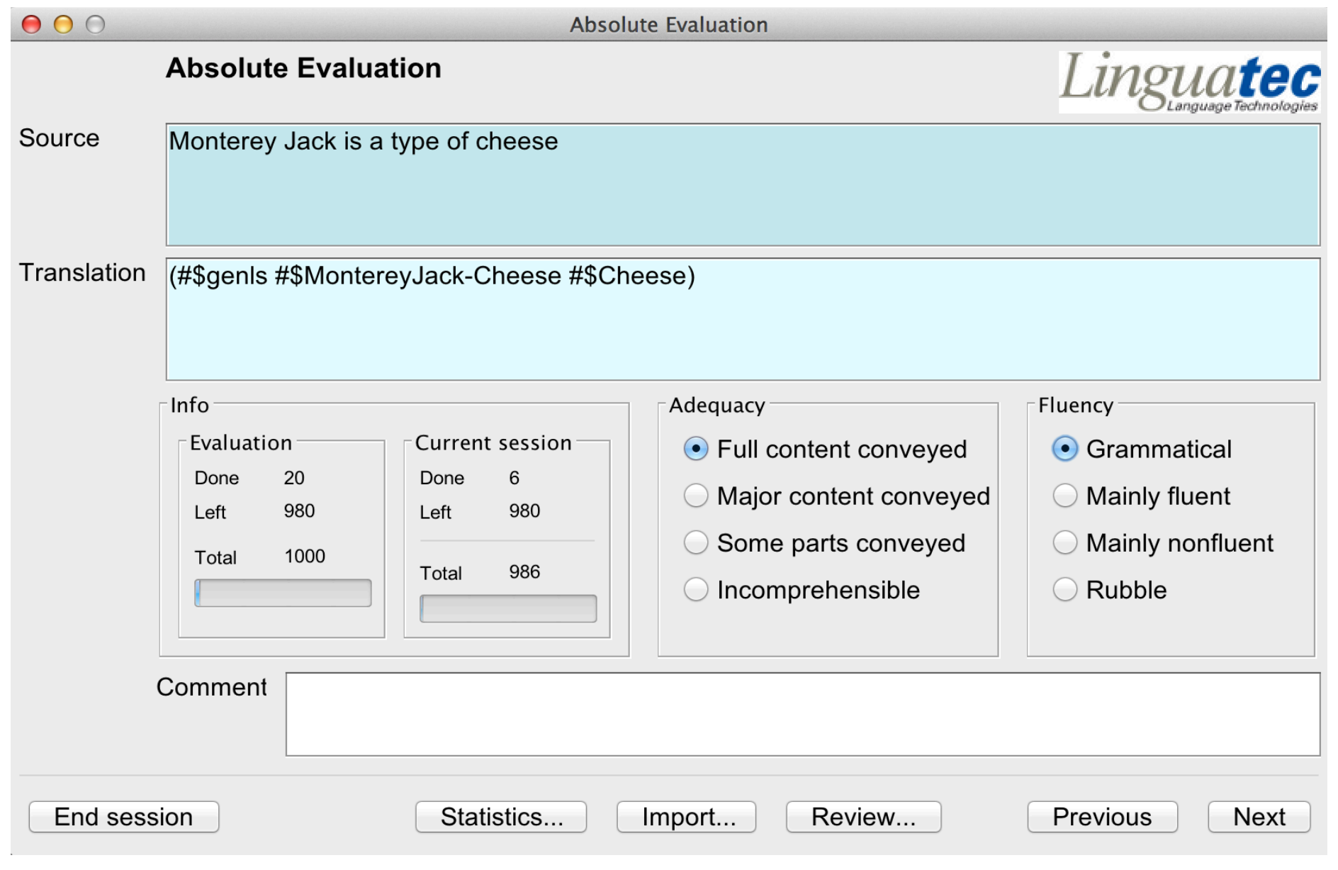


```
<tu>
 <tuv xml:lang="en">
  <seg>Zagreb, Croatia's longitude is 16 degrees</seg>
 </tuv>
 <tuv xml:lang="se">
  <seg>(#$longitude #$CityOfZagrebCroatia (#$Degree-UnitOfAngularMeasure 16.0))</seg>
 </tuv>
</tu>
```

- Using Let’sMT! platform (www.letsmt.eu): Moses
- Evaluation
  - Automatic
  - Human
    - for Adequacy and Fluency
    - using Linguatec’s Sisyphos application for human evaluation
    - 1000 translated “sentences” from evaluation set



| | BLEU Score | NIST Score | TER Score | METEOR Score |
|---|---|---|---|---|
| Case insensitive | 65.26 | 9.1409 | 0.512 | 0.4387 |
| Case sensitive | 54.05 | 7.6859 | 0.6498 | 0.2571 |



Absolute Evaluation

Source: Monterey Jack is a type of cheese

Translation: (#$genls #$MontereyJack-Cheese #$Cheese)

Info
Evaluation: Done 20, Left 980, Total 1000
Current session: Done 6, Left 980, Total 986

Adequacy: ● Full content conveyed ○ Major content conveyed ○ Some parts conveyed ○ Incomprehensible

Fluency: ● Grammatical ○ Mainly fluent ○ Mainly nonfluent ○ Rubble

Comment

- the first results

| Category | Value | Occurences | Percentage |
|---|---|---|---|
| **Adequacy** | Full content conveyed | 209 | 20.9% |
| | Major content conveyed | 289 | 28.9% |
| | Some parts conveyed | 270 | 27.0% |
| | Incomprehensible | 232 | 23.2% |
| **Fluency** | Grammatical | 212 | 21.2% |
| | Mainly fluent | 137 | 13.7% |
| | Mainly non fluent | 244 | 24.4% |
| | Rubble | 407 | 40.7% |

## Project info

**Funded under:** FP7
**Area:** Language Technologies (ICT-2011.4.2)
**Project reference:** 288342
**Total cost:** 4.57M euro
**EU contribution:** 3.55M euro
**Duration:** from Jan 2012 to Dec 2014
**Contract type:** STREP
**Coordinator:** Marko Grobelnik

www.xlike.org

## Project partners

**Institut Jožef Stefan**, Ljubljana, Slovenia
**Karlsruher Institut für Technologie**, Karlsruhe, Germany
**Universitat politecnica de Catalunya**, Barcelona, Spain
**University of Zagreb**, Zagreb, Croatia
**Tsinghua University**, Beijing, China
**Intelligent software components S.A.**, Madrid, Spain
**Slovenska tiskovna agencija d.o.o.**, Ljubljana, Slovenia
**Bloomberg**, New York, USA
**New York Times**, New York, USA (associated partner)
**Indian Institute of Technology**, Mumbai, India (associated partner)

The research within the project XLike leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement nº 288342.